OPEN ACCESS Freely available online

PLOS | COMPUTATIONAL BIOLOGY

Historical Pedigree Reconstruction from Extant Populations Using PArtitioning of RElatives (PREPARE)

Doron Shem-Tov[1]*, **Eran Halperin**[1,2,3]

1 The Balvatnic School of Computer Science, Tel-Aviv University, Tel-Aviv, Israel, **2** International Computer Science Institute, Berkeley, California, United States of America, **3** Molecular Microbiology and Biotechnology Department, Tel-Aviv University, Tel-Aviv, Israel

Abstract

Recent technological improvements in the field of genetic data extraction give rise to the possibility of reconstructing the historical pedigrees of entire populations from the genotypes of individuals living today. Current methods are still not practical for real data scenarios as they have limited accuracy and assume unrealistic assumptions of monogamy and synchronized generations. In order to address these issues, we develop a new method for pedigree reconstruction, *PREPARE*, which is based on formulations of the pedigree reconstruction problem as variants of graph coloring. The new formulation allows us to consider features that were overlooked by previous methods, resulting in a reconstruction of up to 5 generations back in time, with an order of magnitude improvement of false-negatives rates over the state of the art, while keeping a lower level of false positive rates. We demonstrate the accuracy of *PREPARE* compared to previous approaches using simulation studies over a range of population sizes, including inbred and outbred populations, monogamous and polygamous mating patterns, as well as synchronous and asynchronous mating.

Citation: Shem-Tov D, Halperin E (2014) Historical Pedigree Reconstruction from Extant Populations Using PArtitioning of RElatives (PREPARE). PLoS Comput Biol 10(6): e1003610. doi:10.1371/journal.pcbi.1003610

Editor: Alon Keinan, Cornell University, United States of America

Received October 30, 2013; **Accepted** March 13, 2014; **Published** June 19, 2014

Funding: This study was supported in part by a fellowship from the Edmond J. Safra Center for Bioinformatics at Tel-Aviv University. EH is a Faculty Fellow of the Edmond J. Safra Center for Bioinformatics at Tel Aviv University. EH and DST were supported by the German-Israeli Foundation (grant 1094-33.2/2010). URL: http://www.gif.org.il EH was also partially supported by National Science Foundation grant III-1217615. DST was also partially supported by the Israel Science Foundation grant no. 1425/13. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* Email: Email: doronshe@tau.ac.il

This is a *PLOS Computational Biology* Methods article.

Introduction

Pedigree reconstruction is an important problem in the field of computational genetics, with many potential applications such as genealogy inference, heritability estimation, and victim identification [1–4]. Additionally, it has the potential to improve the accuracy of current state-of-the-art relationship inference methods as it uses family structure in a broader sense than just using pairwise genetic similarity information. [5,6]. There are two main variants of the problem, which require different algorithmic approaches. In the first variant, considered by many classical and contemporary papers, the genotypes of several generations are given, and an attempt is made to estimate the pedigree which best explains the observed individuals, as might be the case in wild animal populations. [7–10]. In this paper we consider a more difficult variation of the problem, where we are given the genotypes of the currently living population only, and try to reconstruct the historical pedigree of unobserved ancestors. This variant suits well the scenario of reconstructing the pedigrees of living human populations. [11]. This variant of pedigree reconstruction was previously studied in several theoretical works [12,13]. These papers focus on presenting theoretical bounds on the length of sequence required for reconstructing pedigrees under various combinatorial and stochastic heritability models, but in contrast to our work, do not aim to provide practical solutions for the problem.

The level of difficulty of the problem is highly dependent on the pedigree in consideration. Particularly, small inbred populations pose a considerable challenge since the probability for multiple mating events within any two families is high, and therefore individual pairs usually have more than two last common ancestors (LCAs). Moreover, in small inbred populations there is a complex relationship pedigree graph due to mating within the family.

Recently, three methods tackling pedigree reconstruction from the genotypes of extant individuals were proposed[11,14]; these methods assume monogamy, and synchronized generations. Although unrealistic, these assumptions provide a starting point for developing tools that offer useful methodology. The original paper addressing pedigree reconstruction from the genotypes of extant individuals, presented the methods *COP/CIP* [11]. *COP* assumes infinite population size, and *CIP* tries to reconstruct the pedigree of small inbred populations. *IPED* is a follow-up method, similar in principal to *CIP*, but with improved efficiency [14]. The main idea behind these methods is to construct the pedigree, generation at a time, starting with the given population. In each generation they identify sibling groups using genetic similarity measures, and assign two common parents to each sibling group.

In this work, we point out an important and naturally arising issue of pedigree reconstruction from extant populations, overlooked by all previous methods. We observe that the mother and father of a sibling-group have exactly the same descendants (as

Author Summary

Learning the correct relationships between individuals from genetic data is a basic theoretical problem in the field of genetics, and has many practical consequences. A wide variety of statistical methods for genetic analysis assume the relationships between individuals are known, and can manifest relatedness information to improve inference. The current state-of-the-art methods for relationship inference consider pair-wise genetic similarity, and use it to infer the relationship between each pair of individuals. Reconstructing the pedigrees of an entire population directly has the potential to use more elaborate relationship information, and thus obtains a better prediction of the familial relationships in the population. In contrast to the full set of pair-wise relationships in a population, genetic pedigrees provide a lossless and conflict-free structure for depicting the relationships between individuals. In an effort to make pedigree reconstruction practical we developed a new method, which is an order of magnitude more accurate than previous methods, and is the first method that has the ability to reconstruct polygamous pedigrees.

must be the case for monogamous couples). Since the genotypes of the parents are unobserved, a pairwise relationship analysis relying on the extant descendants will result in maternal relatives having the same likelihood of being related to the mother and to the father, and vice versa (see Fig. 1). Thus, partitioning the relatives into maternal and paternal relatives is required. Undoubtedly, ignoring this issue has a great potential influence on the quality of inferred pedigrees. We discuss a new framework to help understand and correctly deal with this issue, and present a highly efficient algorithm under this framework - *PREPARE* (Pedigree Reconstruction of Extant populations using PArtitioning of RElatives). We extend our method to the case of polygamous pedigrees, and show that our approach results in a considerable improvement in accuracy compared to existing tools, both on monogamous and polygamous pedigrees. Thus, *PREPARE* presents a method that is capable of dealing with more realistic pedigree reconstruction problem as compared to previous methods.

Methods

Similarly to previous methods, we reconstruct the pedigree generation by generation, starting with the last generation, and assuming all of the genotypes of the population come from the same generation. In iteration k, we take the partial $k-1$ generations pedigree, which we call P_{k-1}, and build P_k by adding parents to all of the founder individuals in P_{k-1}. In order to construct the correct pedigree, full-siblings should have two common parents in the pedigree, and half-siblings should have a single common parent. First, we attempt to detect all founder-individual pairs in P_{k-1} which are most likely to be full-siblings, leaving the detection of half-sibling to a later stage. In previous methods, a sibling graph $G=(V,E)$ is constructed, where V includes the set of all founders in P_{k-1}, and E corresponds to the set of pairs of individuals that are likely to be full siblings. Pairs of individuals are considered as potential siblings based on the genetic similarity of the pair's extant descendants. Sibling groups are then detected by finding maximum cliques or proper vertex coloring of the graph G. This approach is problematic, since individuals with equivalent descendant sets, such as parent couples, are completely indistinguishable in the graph G since they have exactly the same set of neighbors. As a result, the siblings graph includes many redundant edges, and fails to represent the true relationship structure.

In contrast with previous methods, we present an alternative graph representation that accounts for the above-mentioned ambiguity, and uses the transitive property of the full-sibling relationship to correctly find the full-sibling groups. We begin each iteration by constructing a contracted siblings graph $G'=(V',E')$. The set of vertices V' is composed of disjoint subsets of V. Particularly, each $v'\in V'$ corresponds to a subset of V, so that for each $v_1,v_2\in v'$ we have $Desc(v_1)=Desc(v_2)$, where $Desc(v)$ represents the set of extent descendants of v (see Fig. 2). Since vertices of G' correspond to subsets of V, we refer to vertices in V' as super-vertices. The set of edges E' corresponds to potential sibling relationship between the corresponding super-vertices, i.e., $(v',u')\in E'$ if there are $v\in v',u\in u'$ such that $(v,u)\in E$. Note that in such case, for every $v\in v',u\in u'$, we will have $(v,u)\in E$. Edges have weights $W\in E'\rightarrow\mathbb{R}$ representing the confidence of the relationship. For a vertex v', we define $contract(v)=v'$ for every $v\in v'$. We provide the details for the construction of the set E' and W in section 2.1.

The key idea of our method lies in a procedure for the assignment of the edges in G' to edges in G in a consistent way. In principle, we are interested in assigning every super-edge $(u',v')\in E'$ to an edge $(u,v)\in E$ that corresponds to the true sibling pair among all pairs in (v',u'). In doing so, we need to take into consideration a set of constraints on the assignments of neighboring super-edges. Ideally, we would like to find the assignment of super-edges to the edges of G, which maximizes the likelihood of the observed population genotypes. In section 2.2, we formulate this problem as an optimization problem using graph terminology, and propose a greedy algorithm which solves it in practice. The assignment algorithm generates an expanded siblings graph $G^*=(V,E^*)$, where $E^*\subseteq E$, denotes the proposed full-sibling pairs, and forms a disjoint clique-cover of the graph.

Under the monogamy assumption, we finish reconstructing the current generation by adding two common-parents to each sibling clique in G^*. In order to account for potential polygamy we add another step that identifies half-siblings and incorporate these into a second graph formulation. Our approach for the reconstruction of polygamous pedigrees relies on two key observations. First, we note that we can treat the full-sibling relation as an equivalence relation, and the half-sibling relation as a relation between equivalence classes. This is true, since if a and b are full siblings, and a and c are half-siblings, then b and c are also half-siblings. According to this observation, we construct a half sibling graph $G_P=(V_P,E_P)$ where V_P corresponds to the equivalence classes defined by the full-sibling relation, and E_P correspond to the half-sibling relation. Second, we observe that the children of every parent in the founder group of P_k correspond to a clique in G_P. We formulate the half-sibling detection problem, as a second graph optimization problem. To solve it, we develop a heuristic algorithm which attempts to find the maximal-weighted set of edges in G_P. The edge set has to satisfy a set of constraints, which represent natural constraints that govern half-sibling relationships.(see section 2.3).

2.1 Constructing the Contracted Sibling Graph

We now describe the construction of the graph $G'=(V',E')$. Recall that the set of super-vertices V' consists of subsets of V that share the same set of extant descendants. For every pair $(v',u')\in V'\times V'$ we have to decide whether $(v',u')\in E'$. In order to do so, we pick a representative pair (v,u), where

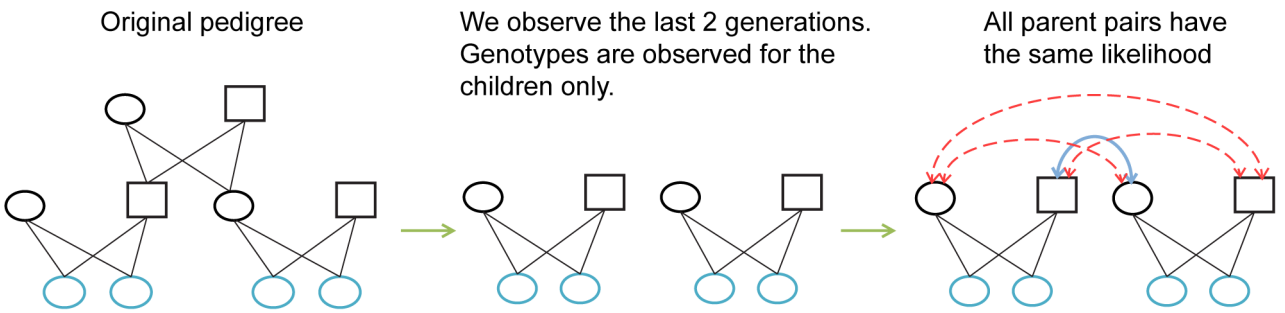

Figure 1. Attempting to reconstruct the simple pedigree on the left, from the genotypes of extant generation (bright blue). Considering observed genetic similarity of extant descendants only, we cannot distinguish which of the four parents in the second generation are siblings (Correctly inferred sibling relationship are colored blue, and wrong potential sibling-relationships in dashed red).
doi:10.1371/journal.pcbi.1003610.g001

$v \in v', u \in u'$, and calculate three scores, corresponding to three putative relations of v and u: unrelated, siblings, and cousins. For each such relationship r, we construct a pedigree $P_r(u,v)$ by adding the relevant ancestry structure. For example, when considering the siblings relationship we construct $P_{siblings}(u,v)$ by adding two common parents for v and u. For unrelated pairs we construct $P_{unrelated}(u,v)$ by adding a different pair of parents to each node (see Fig. 3).

We proceed by simulating inheritance on $P_r(u,v)$; that is, the founders in $P_r(u,v)$ are assigned unique haplotypes and we simulate the recombination process from top to bottom, with a recombination rate of 10^{-8}. We then calculate IBD segments between each pair of extant descendants in $Desc(u)$ and $Desc(v)$ and calculate two *IBD features*: The number of IBD segments, and the total length of IBD sharing (we note that these features of IBD sharing were also considered by *ERSA* [15], a method for the inference of pair-wise family relationships). We repeat these simulations L times for a specified parameter L, thus obtaining an empirical estimate for the distribution of the IBD features. Using the above empirical distributions, we estimate the probability of observing the IBD features for each pair in $Desc(v) \times Desc(u)$ under the relationship r. Since the observed IBD features are typically not observed in any of the L simulations, we use a smoothed form of the distribution using Gaussian kernel smoothing. Formally, let $X_{r1}, \ldots, X_{rL} : X_{ri} \in \mathbb{R}^2$ be the simulated IBD features in the L simulations for a hypothesized relationship r. The density $f_r(x)$ at point x is calculated as:

$$f_r(x) = \frac{|B|^{-\frac{1}{2}}}{\sqrt{2\pi} \cdot L} \sum_{i=1}^{L} e^{-\frac{1}{2}(x - X_{ri})^T \cdot B^{-1} \cdot (x - X_{ri})}$$

Empirical tests led us to the conclusion that scaling the features to have equal variance and using a diagonal bandwidth matrix $B = \beta \cdot I$ with a parameter β in the range 1 to 8 gives the best results. The parameter L compensates running time and accuracy. The accuracy stops improving near $L = 50$, which ends up with a very efficient analysis (See section 2.4 for more details).

Let $IBD_{a,b} \in \mathbb{R}^2$ be the observed IBD features between extant individuals a and b. The above procedure results in a probability $f_r(IBD_{a,b})$, for every $a \in Desc(u), b \in Desc(v)$ and every relationship r in $\{siblings, cousins, unrelated\}$.

For each relationship r, we define

$$score_r(u,v) = \prod_{a \in Desc(u), b \in Desc(v), a \neq b} f_r(IBD_{a,b})$$

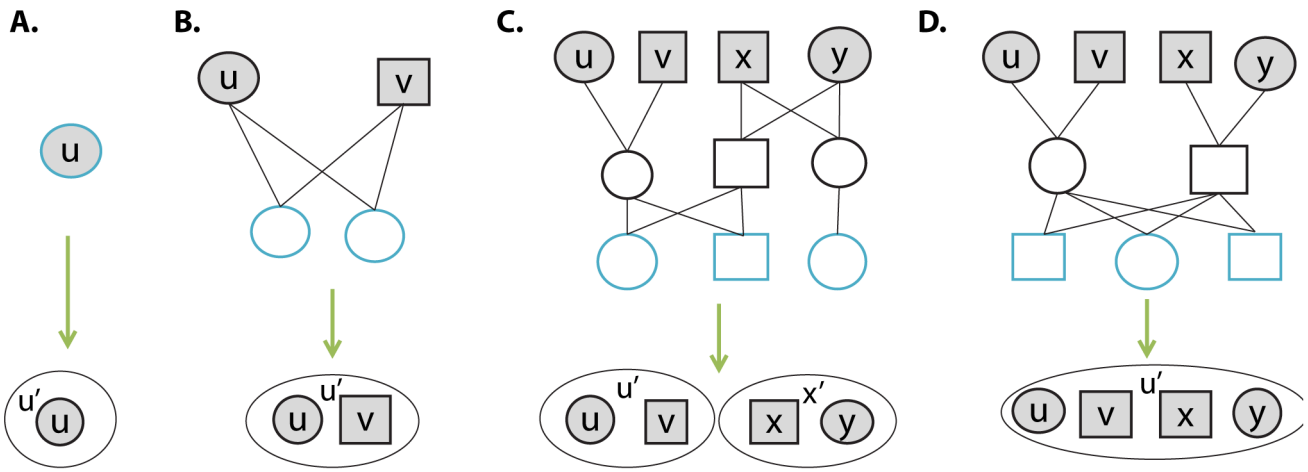

Figure 2. Four examples of vertex contractions, typical for first, second, and third generations. Founders are filled with Grey. Extant individuals are outlined in blue. Green arrows stand for the contraction action.
doi:10.1371/journal.pcbi.1003610.g002

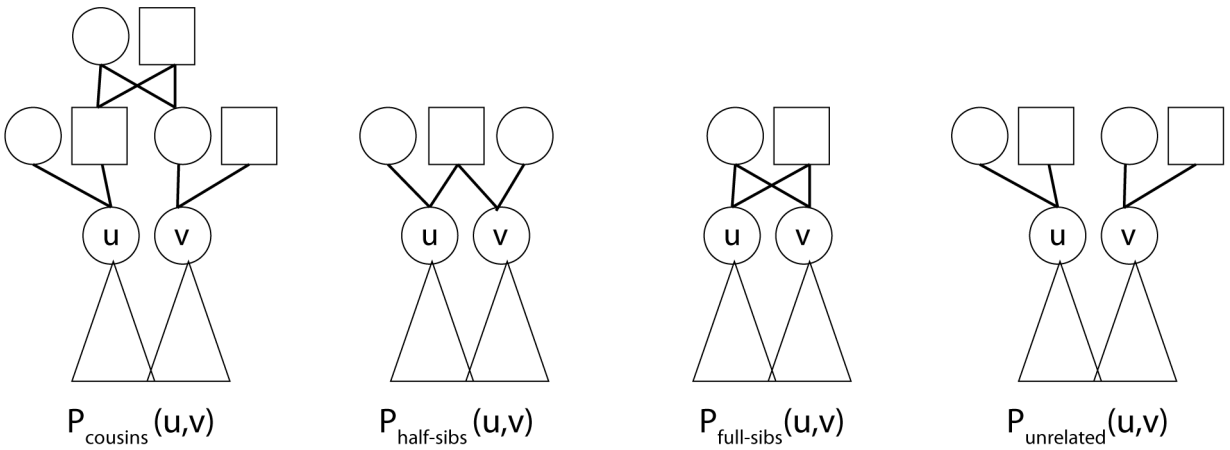

Figure 3. Examples for possible ancestry structures created for individuals u and v in order to test the relationship between them. The triangles under u and v represent their existing descendants, edges represent parent-offspring relationship.
doi:10.1371/journal.pcbi.1003610.g003

We note that $score_r(u,v)$ can be intuitively interpreted as a composite likelihood of r. If $score_{siblings}(u,v)$ is larger than $score_{unrelated}(u,v)$ and $score_{cousins}(u,v)$ we add (u',v') to E' with the weight

$$w(u',v') = \frac{score_{siblings}(u,v)}{\sum_r score_r(u,v)}$$

Fig. 4 shows the distribution of $\frac{score_r(u,v)}{\sum_r score_r(u,v)}$ under different true relationships. Notice that cases where u,v are distantly related (cousins, 2nd-cousins etc.) will tend to have a maximal score under $score_{cousins}(u,v)$. This is desirable, since we only seek to distinguish siblings from non-siblings at this point.

2.2 The Assignment Algorithm

In the assignment stage, we are given the contracted siblings graph $G'=(V',E')$, and we search for an assignment of a sibling relation between super-vertices, depicted by an edge $(u',v')\in E'$ to a single sibling-relation between two individuals $(u,v)\in E$. Our assignment needs to obey the transitivity constraint of the full sibling relation. Recall that the weight of an edge $w(u',v')$ corresponds to the strength of evidence for the existence of a sibling pair (u,v), where $u\in u', v\in v'$. We therefore formulate the edge assignment problem as follows:

***Problem 1.* Maximum weight disjoint clique cover edge assignment.** Given the contracted graph $G=(V',E')$, find the maximal-weight set of edges $E^*\subseteq E$, such that E^* is a legal assignment of E', under the constraint that the set of assigned edges E^* forms a clique cover of the graph $G=(V,E)$, i.e., E^* is composed of an edge-disjoint set of cliques.

We first show that the above problem is NP-hard:

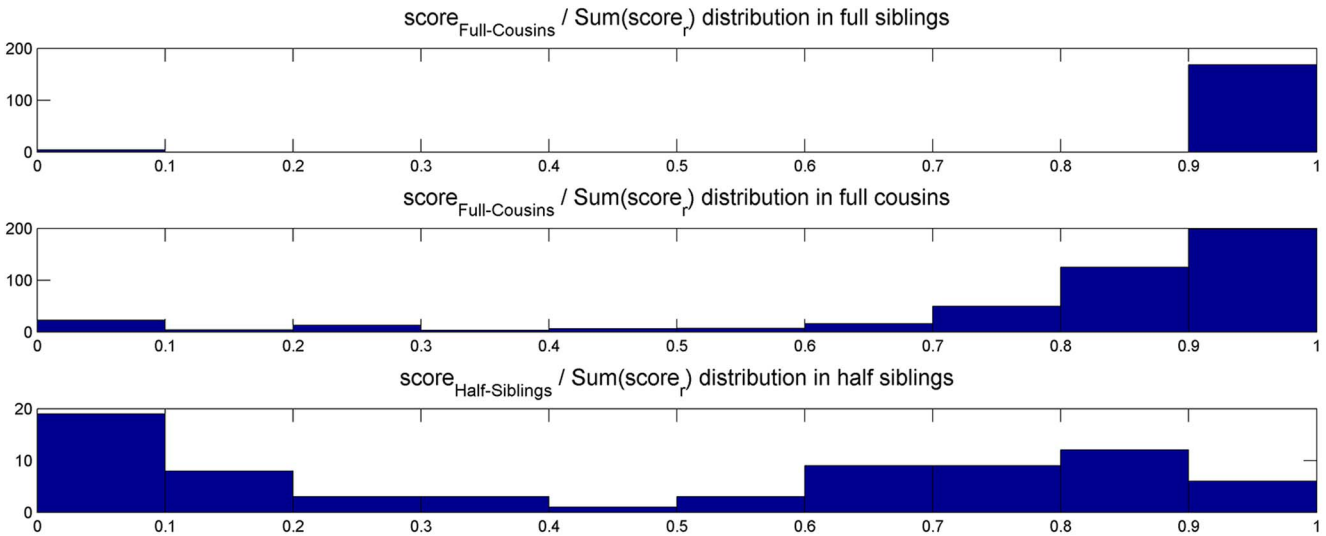

Figure 4. Distribution of relationship scores under specific true relationships.
doi:10.1371/journal.pcbi.1003610.g004

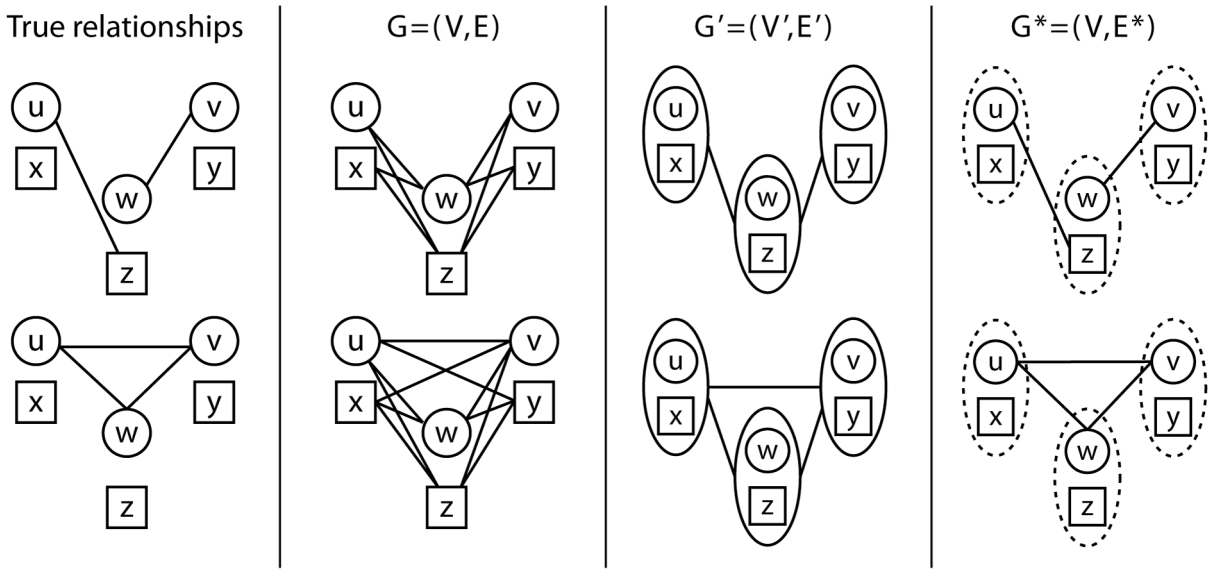

Figure 5. Intuition for sibling assignment, depicting the potential-siblings graph G, the contracted graph G', and assigned graph G^*. In both examples $(u,x),(v,y),(w,z)$ are parent couples with extant descendants in the observed population. A. For the case where $(u,z),(w,v)$ are full-siblings, the contraction will end in G' composed of three super-vertices, connected by two edges; the assignment algorithm will assign each edge to a disjoint clique. B. If (u,v) are also full-siblings, a 3-clique is formed in G'; the assignment algorithm assigns all edges to a corresponding 3-clique of siblings.
doi:10.1371/journal.pcbi.1003610.g005

Theorem 1. *The maximum weight disjoint clique cover edge assignment is NP-hard.*

Proof. We will show a reduction from maximum clique. In [16] it is shown that it is NP-hard to decide whether a graph $G=(V,E)$ has a clique of size $n^{1-\epsilon}$ or if its largest clique is smaller than n, where $=0.01$. Consider an instance $G=(V,E)$ to the clique problem, and let C be its largest clique. We define $G'=(V',E')$, where $V'=V$, and $E'=E$. Thus, any clique cover of G is a legal assignment of G'. Note that if $|C|>n^{1-\epsilon}$ then the optimal clique cover is necessarily of size at least $\frac{n^{2-2\epsilon}}{2}$. On the other hand, if $|C|<n^{\epsilon}$ then it is easy to see that the optimal clique cover is obtained in case all cliques in the cover are of size n^{ϵ}, and thus the clique cover size is of size at most $\frac{n^{1+\epsilon}}{2}$. Thus, if the Maximum Weight Disjoint Clique Cover Edge Assignment was polynomial, then we could decide in polynomial time between the case where the maximum clique is of size n and the case where the maximum clique is of size $n^{1-\epsilon}$, which is an NP-hard problem.

We therefore apply the following greedy algorithm. We will need to introduce a few notations. First, we treat vertices $v'\in V'$ as vertices in G', as well as subsets of V, depending on the context. For each $x\in V$, we denote by $N_{E^*}(x)$ the set of neighbors of x in E^*. Moreover, we define $N_{E'}(x)=\{contract(y)|y\in N_{E^*}(x)\}$, i.e., the set of super-vertices corresponding to the neighbors of x in E^*. Finally, let $N_0=\{x\in V|\ |N_{E^*}(x)|\ =0\}$.

We start by setting $E^*=\varnothing$. The algorithm proceeds by traversing all super-edges $(u',v')\in E'$ in decreasing weight order. In each iteration the set E^* consists of a set of disjoint cliques of G, and E' consists of a set of yet to be assigned edges. For each $v\in N_0\cap v'$ and $u\in u'$ we say that v can be added to the clique of u if for every $x'\in N_{E'}(u)$ we have that $(x',v')\in E'$. Similarly, we say that $u\in N_0\cap u'$ can be added to the clique of $v\in V$ if for every $x'\in N_{E'}(v)$ we have $(x',u')\in E'$. When traversing an edge (u',v') we search for a pair (u,v) where u has the maximal clique size, $|N_{E^*}(u)+1|$, from within u', $v\in N_0\cap v'$, and v can be added to the clique of u (or in a symmetric manner that u can be added to the clique of v and $|N_{E^*}(v)+1|$ is maximized). We then assign (u',v') to (u,v) by

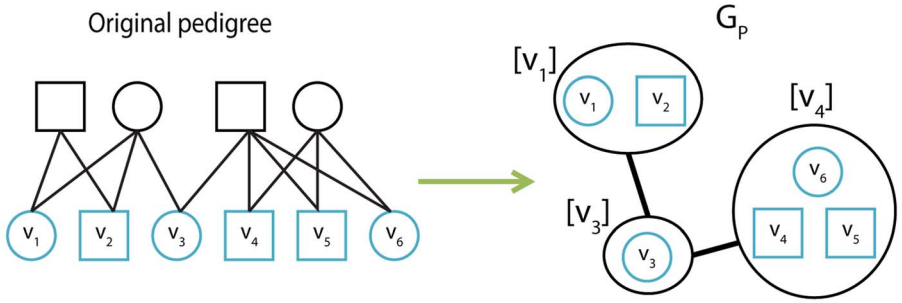

Figure 6. An example for the construction of G_P in the first generation.
doi:10.1371/journal.pcbi.1003610.g006

Table 1. Sensitivity and PPV scores (as defined in the results section) of half-siblings using two coloring order schemes.

	PREPARE		naive	
Population size	**Sensitivity**	**PPV**	**Sensitivity**	**PPV**
200	1.0	0.91	0.91	0.91
300	0.91	0.88	0.79	0.85
400	0.97	0.88	0.85	0.88
500	1.0	0.88	0.88	0.91

(1) PREPARE's greedy coloring scheme as described in section 2.3. (2) Coloring cliques from the heaviest to lightest; if possible color with F_P, else if possible color with M_P.
doi:10.1371/journal.pcbi.1003610.t001

adding (u,v) to E^*, and removing (u',v') from E'. We also assign (x',v') to (x,v) for every $x \in N_{E'}(u)$.

Fig. 5 summarizes the contraction and assignment stages with an example. Note that cases such as 3-cliques in G' (Fig. 5-B) can have multiple assignments with the same score (3 siblings from one parent couple, or 3 pairs of siblings from 3 different parent couples). In such cases our algorithm chooses the more parsimonious solution in which there is a smaller number of parents.

2.3 Half-sibling Detection

In the following stage we define the half-sibling detection problem, where we attempt to detect groups of individuals with a single common-parent. First, we define the full-sibling relation, on individuals: $FS = \{(u,v) | u,v \textit{ have two common parents}\}$. Notice that FS is defined as being reflective, and thus it is an equivalence relation on V. V/FS is the quotient set of V on FS, which in this case is simply the set of disjoint groups of full-siblings. We obtain FS from the edges in E^* computed in section 2.2. E^* is a clique cover, and so naturally describes an equivalence relation.

We define $HS = \{([u]_{FS}, [v]_{FS}) \mid \forall (u,v) \in [u]_{FS} \times [v]_{FS}, u, v \textit{ share exactly one common parent}\}$, which is the half-sibling relation, as a relation between equivalence classes in V, in respect to FS. Assuming the pedigree is known, HS is defined properly since if u and v are full siblings, and u and x are half-siblings, than v and x are half siblings. This allows us to simplify the half-sib detection problem, by constructing the polygamy graph $G_P = (V_P, E_P)$, where $V_P = V/FS$ s.t each vertex $v \in V_P$, represents a group of full-siblings, and each edge $(u,v) \in E_P$ represents a half-sibling relation between u and v (see Fig. 6). The edges are added to E_P, with a similar stage to 2.1, only the hypotheses tested this time are made for siblings groups $(u,v) \in E_P$, and are relevant to the half-sibling case (half-siblings,cousins,unrelated).

The graph G_P has the convenient property that if a group of individuals $\{v_1,...,v_k\} \subseteq V$ have a single-common-parent then $\{[v_1],...[v_k]\} \subseteq V_P$ form a clique in G_P. We thus assume by parsimony, that each clique $\{u_1,...u_k\}$ in G_P connects all of the children of a single parent w, such that each $u_i \in \{u_1,...u_k\}$ is a full-sibling-group which contains the children of w and a single mate. We therefore formulate the half-sib detection problem, as follows:

Problem 2. **Maximum weight, two-color clique cover.** Given the graph $G_P = (V_P, E_P)$, find sets of edges $F_P, M_P \subseteq E_P$, such that both F_P and M_P consist of an edge-disjoint set of cliques, $F_P \cap M_P = \varnothing$, and the total weight of F_P and M_P is maximized.

Theorem 2. *The Maximum Weight Two Color Clique Cover is NP-hard.*

Proof. We will show a reduction from maximum clique. Consider an instance $G = (V,E)$ to the clique problem, and let C be its largest clique. If $|C| > n^{\epsilon}$ we can set $M = C$ and $F = \varnothing$, and therefore the optimal solution to the coloring problem has at least $\frac{n^{2-2\epsilon}}{2}$ edges. On the other hand, if $|C| < n^{\epsilon}$ then the size of each of

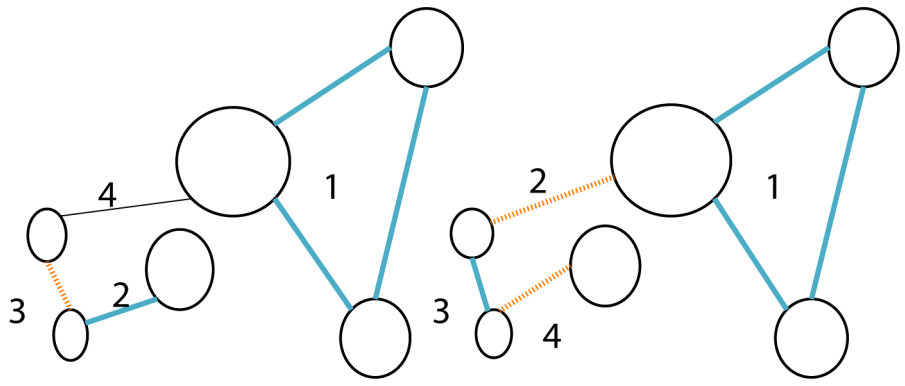

Figure 7. An example for a case where the coloring order we purpose enables coloring more cliques with two colors than coloring the same graph with an arbitrary order. The coloring order is depicted near the cliques. In the left graph we follow the depicted order and color the clique blue if possible, else we color it dashed-orange. The fourth click cannot be colored since it touches a blue and a dashed-orange clique. In the right graph we use our coloring scheme, which prefers coloring cliques touching cliques that are already colored. Using this order we are able to color all four cliques.
doi:10.1371/journal.pcbi.1003610.g007

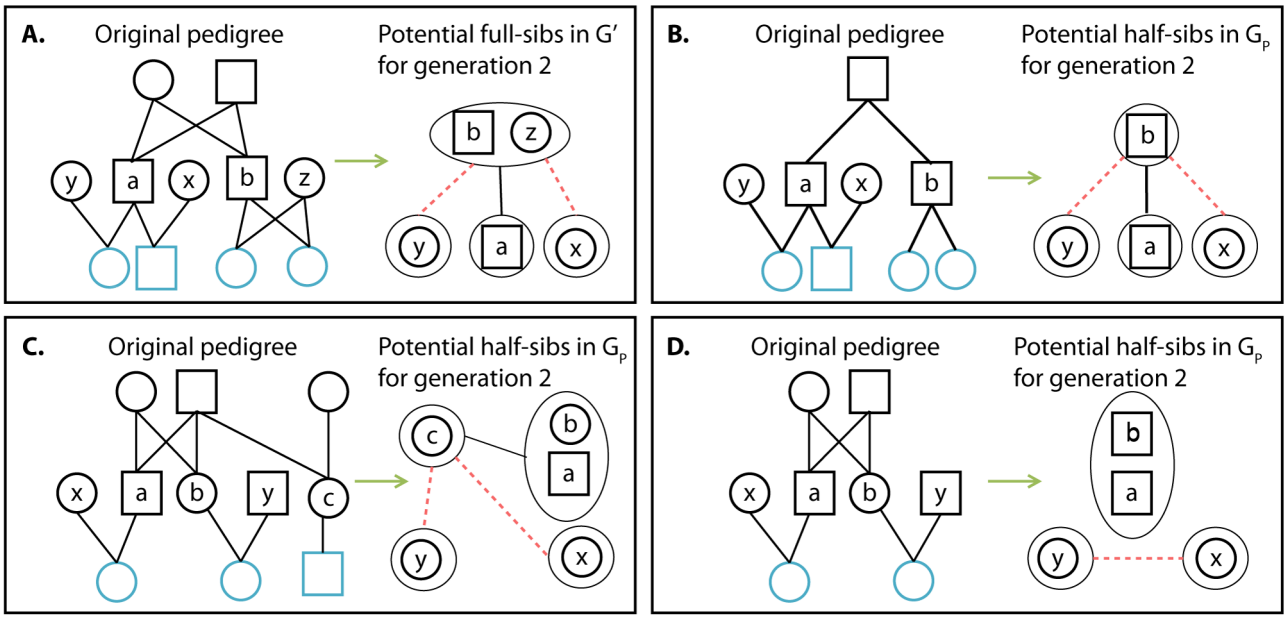

Figure 8. Depicting cases where edge removal rules are required in polygamous pedigree reconstruction. Redundant graph edges are dashed red, correct edges in solid black.
doi:10.1371/journal.pcbi.1003610.g008

M and P is at most $\frac{n^{1+\epsilon}}{2}$, and thus the total size of both of them is bounded by $n^{1+\epsilon} < n^{2-2\epsilon}/2$. Thus, by solving the Maximum Weight Two Color Clique Cover in polynomial time we can decide between graphs with clique size at most n^{ϵ} and graphs with clique size at least $n^{1-\epsilon}$, hence the problem is NP-hard.

Informally, we try to color all edges E_P in two colors, F_P and M_P, s.t each color creates a set of disjoint cliques. F_P colored cliques, represent full-sibling-group cliques with a single common father, and M_P colored cliques, represent full-sibling-group cliques with a single common mother.

This problem is also NP-hard and we therefore use the following greedy approach. For simplicity, we assume G_P is connected. The algorithm begins by setting $F_P = M_P = \varnothing$. We will denote by $V(F_P)$ and $V(M_P)$ the set of vertices induced by F_P and M_P respectively. The algorithm proceeds in iterations. In each iteration we search for the heaviest clique $C \subseteq V_P \backslash V(F_P)$ such that $C \cap V(M_P) \neq \varnothing$, and the heaviest clique $C \subseteq V_P \backslash V(M_P)$ such that $C \cap V(F_P) \neq \varnothing$. Without loss of generality, assume that the heaviest among those is a clique C in $V_P \backslash V(F_P)$. If C contains only one vertex, we search instead for the heaviest clique C in $V_P \backslash (V(F_P) \cup V(M_P))$. We add the edges of C to F_P and remove these edges from the graph. Clearly, both F_P, M_P consist of a set of disjoint cliques of G_P.

Notice that we try to minimize the number of arbitrarily colored cliques, by choosing cliques adjacent to cliques that are already colored. Simulation studies show that choosing this coloring order increases the half-sibling sensitivity from 85% to 97% on average (see table 1). It is easy to see that sub-graphs that are composed of a connected list of cliques will be colored optimally by our coloring scheme. An example for such a graph is depicted in Fig. 7.

The graph formulation of the half-sibling detection assumes that each edge in E_P represents a unique half-sibling relationships. We notice, that in some cases G_P might contain redundant edges. In order to simplify the explanation, we extend the definition of $Desc(u)$ to nodes in G_P: $Desc_P(u) = \bigcup_{v \in [u]} \{x \mid x \in Desc(v)\}$. The problem arises, when there exists a pair of nodes u,v from the same generation, such that $Desc_P(u) \subseteq Desc_P(v)$. In such a case, an edge (u,x) may be added to E_P, as a result of a relationship (v,x). Trying to contract u and v is not sound, since different relationships can be detected for u, and v to a third vertex x, by testing them separately. Instead, we apply a preprocessing to G_P, in the form of a set of parsimonious rules. The rules aim at filtering

Table 2. Running times of PREPARE on 1.6GHz Intel Core i5-2467M machine with 4G RAM using a single thread.

Population Size	monogamous	polygamous
100	31s	4m 18s
200	53s	9m 21s
500	4m 55s	56m 40s
1000	10m 27s	93m 41s

The two parameters affecting the running time of prepare is the population size, and whether PREPARE is run on monogamous or polygamous mode. Most of the running time is spent on reconstructing the fifth generation.
doi:10.1371/journal.pcbi.1003610.t002

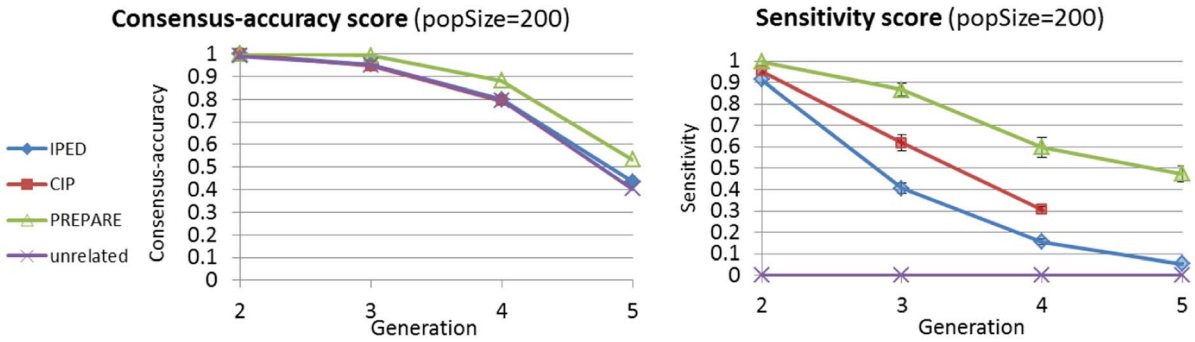

Figure 9. Example for the problematic nature of the consensus-accuracy score, in contrast with the sensitivity score we propose. Notice how the unrelated pedigree structure receives similar consensus-accuracy scores to *IPED* and *CIP* reconstructions. Still, *PREPARE* scores are significantly higher. Shown are average scores over 5 simulations, and standard deviation bars. (Some error bars are too small to be visible).
doi:10.1371/journal.pcbi.1003610.g009

all the edges, except the ones that explain the observed features in the simplest way.

The first rule we apply concerns the case depicted in Fig. 8-A. In this case, an individual a, with a half-sibling b, has children with two mates x, and y. Since a,b,x,y do not have full siblings, each of them is represented in G_P as a sibling-group of one individual. Since x and y have children only with a, their descendant sets are contained in a's descendant set. As a result, half-sibling edges should form between (x,b) and (y,b), additionally to the correct edge (a,b). To deal with this case, if we find a node a, in V_P that has two mates, x,y and the following holds: $(x,b)\in E_P$ *AND* $(a,b)\in E_P$, we remove (x,b) (we do the same for (y,b)). A similar rule is applied to the contracted graph G', where redundant full-sibling edges result from an equivalent case to the one just mentioned, and are removed in the same manner (see Fig. 8-B). A third rule is applied to G_P to deal with a case similar to the one in rule 1, only the mates x,y are not the mates of a single individual a, but instead x is the mate of a, y is the mate of b and a,b are full-siblings (see Fig. 8-C). In such a case, a true relation (b,c) may cause redundant half-sibling edges $(x,c),(y,c)$. These cases are characterized by mates x,y that have few or no full-siblings. Thus, we look for edges $(a,c),(x,c)$ where $|[x]_{FS}| < |[a]_{FS}|$, such that x is the mate of a, and remove (x,c) from G_P. Finally, we observed half-sibling edges forming between two mates (x,y), of (a,b) such that (a,b) are full-siblings. This results from the fact that most of a and b's descendant similarity was already explained by the formation of the full-sibling relationship (a,b). The difference between the half-sibling hypothesis and the null hypothesis for (x,y) becomes small. As a result, noisy decisions are made. To handle this final case, we remove half-sibling edges between mates of full siblings (a,b) if they have a half-sibling edge (x,y) in G_P(see Fig. 8-D).

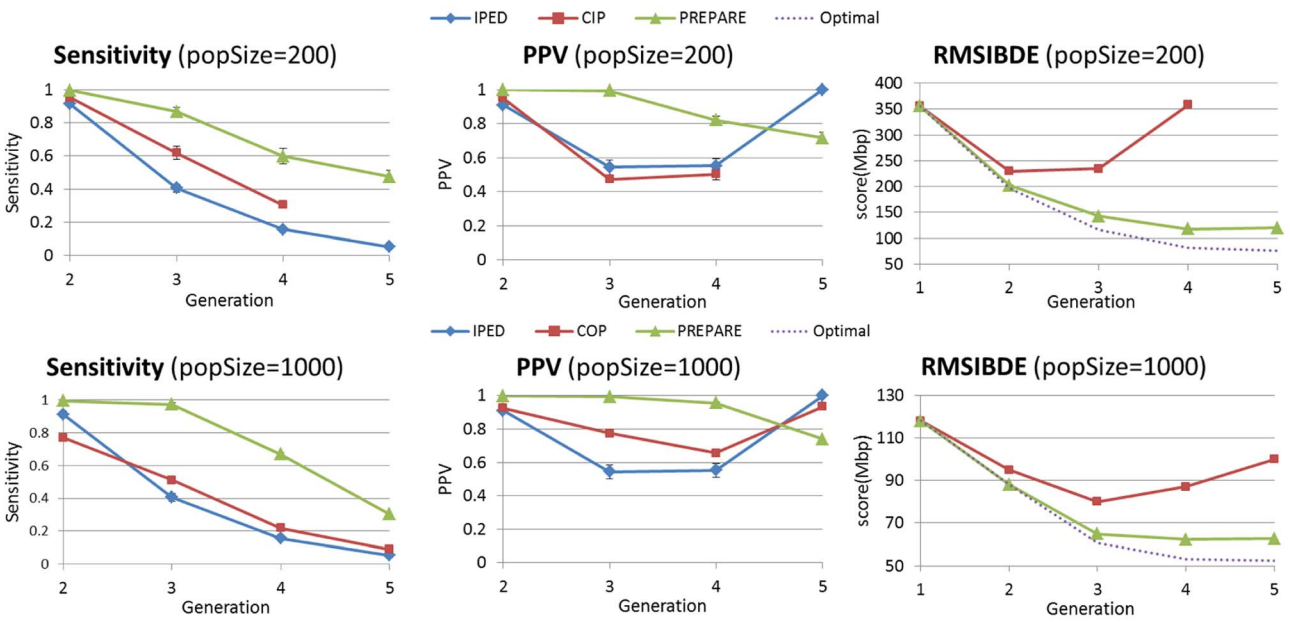

Figure 10. Comparison of pedigree reconstruction methods for monogamous populations, using Sensitivity, PPV, and *RMSIBDE*. Populations were simulated with Wright-Fisher simulations of 5 generation. Shown are average scores over 5 simulation, with standard deviations bars. The optimal *RMSIBDE* score is calculated by scoring the true k-generation pedigree. The first generation pedigree in the *RMSIBDE* figures, is the score of the pedigree where all individuals are unrelated, and is shown as reference. (Some error bars are too small to be visible).
doi:10.1371/journal.pcbi.1003610.g010

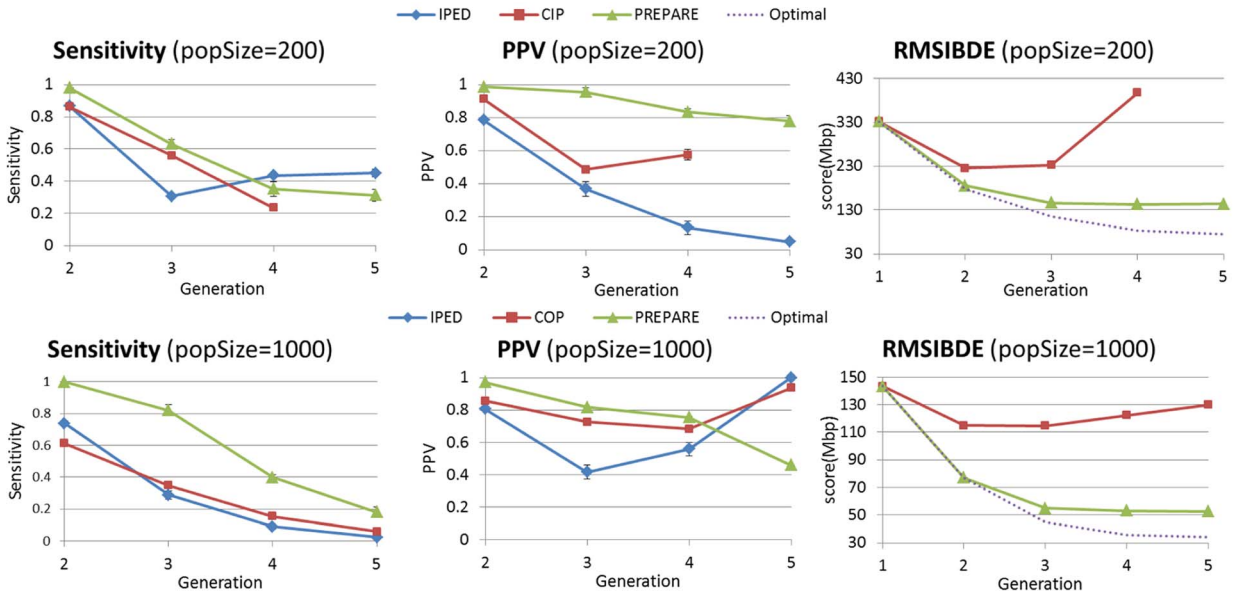

Figure 11. Comparison of pedigree reconstruction methods for polygamous populations. Populations were simulated with polygamous Wright-Fisher simulations of 5 generation. Shown are average scores over 5 simulation, with standard deviations bars. (Some error bars are too small to be visible).
doi:10.1371/journal.pcbi.1003610.g011

2.4 Efficiency Considerations

Simulating inheritance for the descendants of every two individuals during the graph constructions is very time consuming, and is the reason *CIP* is impractical for large populations, or pedigrees deeper than 4 generations. Notice that if a pair of extant descendants has exactly the same ancestor structure in the pedigree, than the simulated IBD features are sampled from the same distribution. *IPED* purposes caching individual pairs with identical inheritance paths, and introduces an accompanying dynamic programming algorithm for minimizing the number of operations.

In *PREPARE*, we use a simplified version of this idea. For every pair (a,b) of extant descendants, we calculate a least-common-ancestors (LCAs) vector $LCAs(a,b)$, which is a list of the meiosis distances between (a,b) and their least common ancestors. For example, all full-siblings will have the $LCAs(a,b) = [1,1]$, since full-siblings always have two common ancestors, with one separating meiosis. We hash the simulated distribution for this LCA vector, where the key represents the vector itself, and the value is the distribution. We simulate inheritance only when needed, i.e. when $u',v' \in V'$ have at least one descendant pair, without a hashed distribution, thus saving most of the redundant computation. Practically, the running time of *PREPARE* is equivalent to the running time of *IPED*, and is even slightly faster (see Table. 2). Although $LCAs(a,b)$ does not capture completely the ancestry structure for (a,b), we observed empirically (data not shown) that running simulations for each ancestry structure does not improve the reconstruction accuracy. Apparently, pairs of individuals (a,b) with the same LCAs vector have similar IBD distributions. The similarity is large enough to make the repetition of inheritance simulation for two such pairs redundant.

2.5 Availability

The *PREPARE* method, inheritance simulators, and quality evaluation tools are available at http://www.cs.tau.ac.il/~heran/cozygene/software.shtml

Results

We compare the accuracy of our method to previous pedigree reconstruction methods on numerous simulations. Different simulations include combinations of population size and inheritance modes (monogamous and polygamous). Smaller population sizes correspond to inbred populations with multiple relationships between families. Larger populations correspond to outbred populations, with simpler pedigree structures. We also study the effect of population bottlenecks on the reconstruction quality. In order to test *PREPARE* on a more realistic scenario, we run it on a realistic simulation starting from HapMap phaseIII *CEU* and *YRI* populations as founders. The simulation simulates polygamous random mating in this population for 200 years, reaching to a final population size of 1000. Finally, we apply PREPARE on the HapMap *MEX* population as a feasibility test for application of our method for real populations.

3.1 Simulations

Similarly to previous methods, we use a Wright-Fisher (WF) simulator that includes recombination and genders. We add several new features, which makes this simulator more flexible. First, we add the ability to control polygamy through a polygamy probability parameter p, which controls the probability for an individual to have a child with more than one mate. Second, we add an option to simulate dynamic population sizes by specifying an initial population size and a final population size. The simulator calculates the required population change per generation and modifies the population size with that ratio in every generation.

Additionally, we experiment with a more realistic forward simulator that does not assume synchronized generations, and allows polygamy. We simulate inheritance as a function of time, where individuals can have children after the age of 20, and die at an age drawn from a capped exponential distribution with mean 50. The birthrate is changed according to the current population size, and is tuned to reach a predefined target population size. This simulator produces actual recombined haplotypes, from the

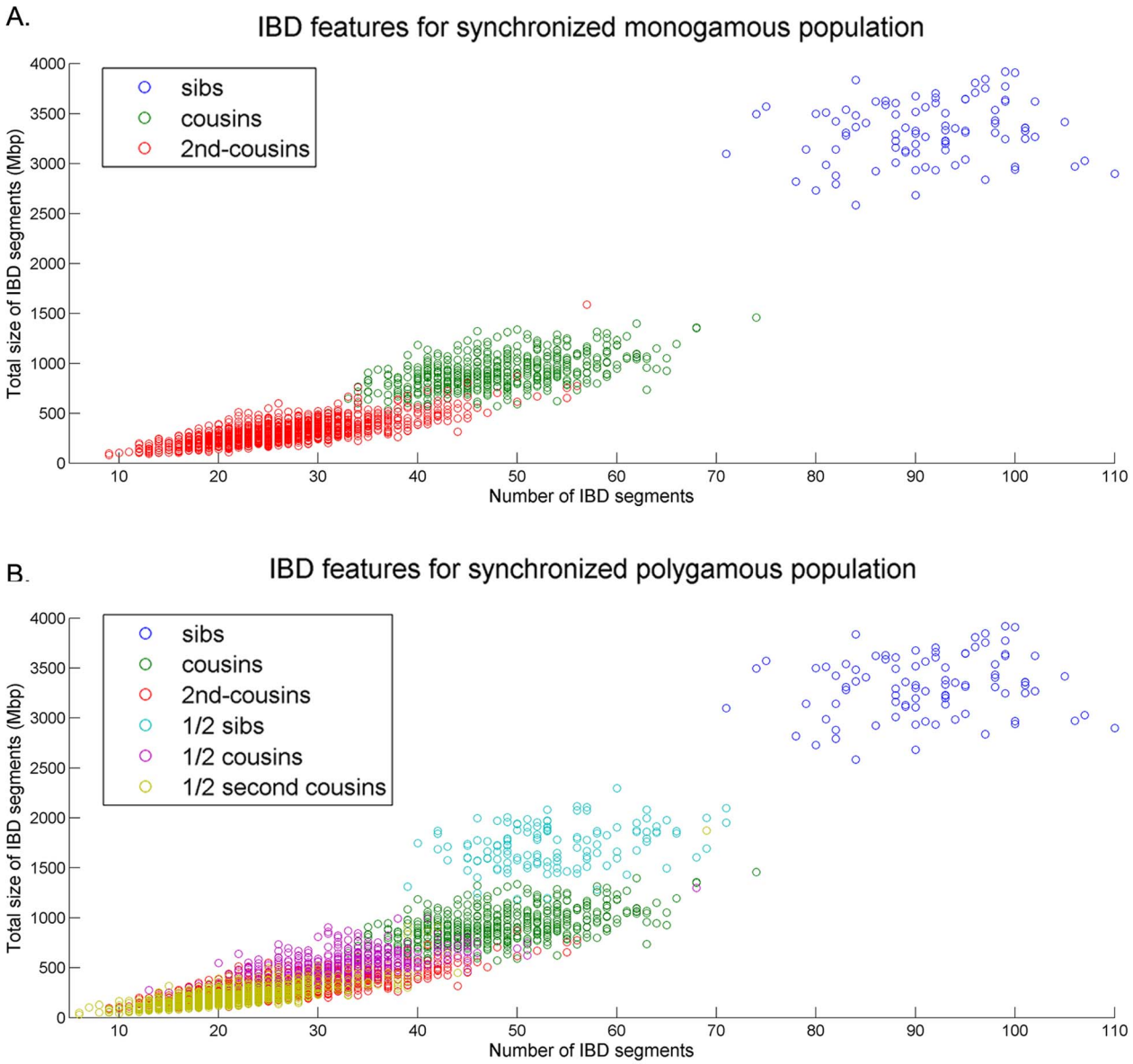

Figure 12. Simulated IBD feature distribution in monogamous and polygamous populations. The overlap in polygamous distributions is the main challenge in reconstructing pedigrees of real populations.
doi:10.1371/journal.pcbi.1003610.g012

haplotypes of 160 *CEU* and *YRI* HapMap representatives. More specifically, the simulation runs in 5 year iterations, and a pool of unmated mature individuals is maintained at all times. Every iteration, individuals from the pool are matched to uniformly drawn mates. A matching has probability m_p to succeed. Every mated pair has a probability p_b to have a child, where p_b is initialized to be 1, and is modified in every iteration by +0.2 or -0.2 depending on whether the current population size is smaller or larger than the target population size. Polygamy is achieved through second-marriage, which can occur since once a mate dies, the individual is added back to the unmated pool. Finally, in order to include possible IBD detection errors, we detect IBD segments from simulated genotypes using *GERMLINE*, [17], and extract the IBD-features information from its output. This simulator also has the advantage of having a possible dynamic population size. The population grows or shrinks depending on the initial and target population sizes.

3.2 Quality Evaluation

Many different measures can be accounted in evaluating the quality of reconstructed pedigrees. We first use a previously defined score, to compare *PREPARE* to previous methods. For the large part of the presentation, we define and use other natural evaluation scores, which we deem as more relevant, and interpretable. In previous methods, a consensus-accuracy score, which counts the number of extant individual-pairs with the same minimal meiosis-distance as in the true pedigree was used [14]. This score treats correct detection of unrelated pairs and related pairs identically. This is problematic since the number of unrelated pairs dominates the score. For example, a trivial algorithm that outputs a pedigree where all individuals are unrelated receives a

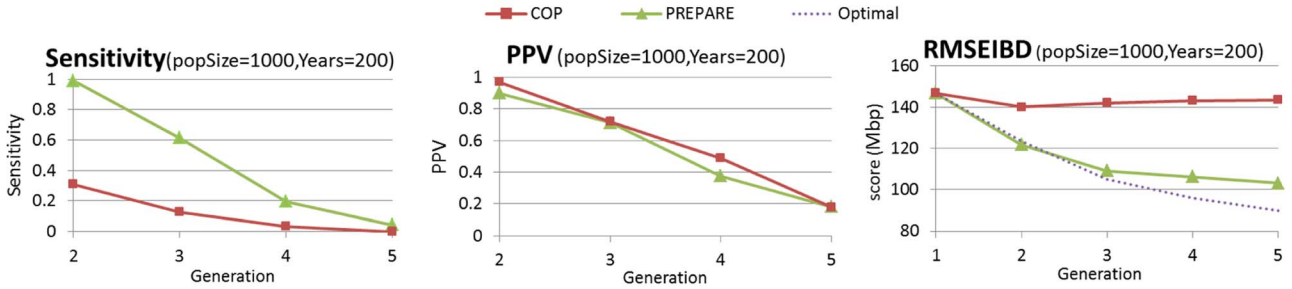

Figure 13. The performance of PREPARE on realistic simulation is comparable to polygamous Wright-Fisher simulations. The simulated population grew from 160 individuals of the *CEU* and *Y RI* HapMap populations to 846 individuals in 200 years. This simulation accounts for IBD detection errors, asynchronous mating and dynamic population size.
doi:10.1371/journal.pcbi.1003610.g013

high consensus-accuracy score (see Fig. 9). As a new standard for pedigree-reconstruction evaluation, we suggest three types of scores: sensitivity, positive-predictive-value (PPV), and IBD-length prediction error.

We define sensitivity as the fraction of correctly constructed (distance wise) related pairs from the total number of related pairs in the original pedigree. PPV is defined as the fraction of correctly constructed related pairs from the total number of related pairs in the reconstructed pedigree. More formally, define R as the reconstructed pedigree, O as the original pedigree, $D_P(i,j)$ as the minimal number of meiosis separating i and j in pedigree P, and $RelExtant_P$ as the set of extant-individuals, which are related according to pedigree P. Let $TP_{R,O} = \sum_{i,j \in RelExtnat_O} I_{(D_R(i,j) = D_O(i,j))}$. Then,

$$Sensitivity_{R,O} = \frac{TP_{R,O}}{|RelExtant_O|}\ ,\ PPV_{R,O} = \frac{TP_{R,O}}{|RelExtant_R|}$$

We run *PREPARE* for G generation, and compare the scores of reconstructed pedigrees for every generation $k \in \{1 \ldots G\}$ against the first k generations of the original pedigree. This way we can assess the accuracy of different relatedness degrees ($k = 2$ corresponds to siblings, $k = 3$ to siblings and first-cousins, etc.)

Scores such as sensitivity and PPV have the disadvantage of not weighing mistakes according to their magnitude. A second disadvantage is that the minimal meiotic distance does not capture the full complexity of a real pedigree (for example, double cousins detected as cousins will get a full scoring). For these reasons, we suggest to alternatively measure pedigree quality by calculating the root mean square IBD-length error (*RSMIBDE*):

$$RMSIBDE = \sqrt{\sum_{i,j \in Extant} (IBD_R(i,j) - IBD_O(i,j))^2}\ ,$$

where *Extant* is the set of extant individuals in the population, IBD_O is the observed total length of IBD segments between individuals i and j, and IBD_R is the total length of IBD segments between individuals i and j, as given from simulating inheritance on the reconstructed pedigree R. Since this score is dependent on the randomized scoring-simulation, we average the score of 5 runs. The *RSMIBDE* can be interpreted as the expected prediction error (in Mbp) of the typical pair-wise total-IBD-length, given the reconstructed pedigree.

3.3 Comparing *PREPARE* and Competing Methods on Monogamous Simulations

We tested the competing methods on monogamous Wright-Fisher simulated population, of constant sizes: 100, 200, 500, and 1000. When it was possible, we ran *CIP* (up to 4 generations due to its high runtime complexity), and for larger populations we ran *COP*. *PREPARE* was run in monogamous mode. Results on 100 and 200 individuals were similar, as well as results for 500 and 1000 individuals. In Fig. 10, we compare the three methods for small populations (200) and larger populations (1000). In all the scenarios we tested, *PREPARE* was the most sensitive; for pedigrees of up to 5 generations (corresponding to 3rd cousins) and populations as small as 100 individuals. For the larger populations, the improvement in sensitivity is highest, where *PREPARE* is able to build a pedigree which correctly predicts the minimal meiosis

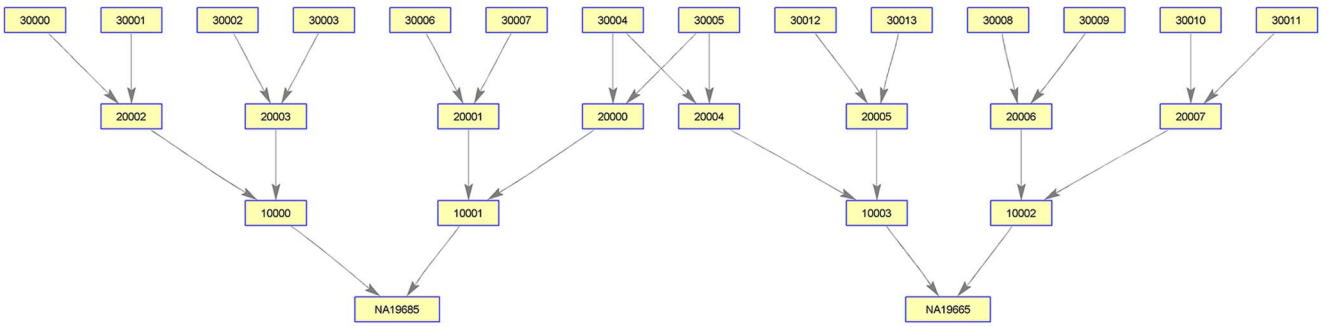

Figure 14. PREPARE successfully isolates the 4 generation pedigree found by CARROT. Nodes correspond to individuals, and edges to parent offspring relationships. The last generation individuals are real HapMap individuals, and the other nodes are ancestors predicted by PREPARE.
doi:10.1371/journal.pcbi.1003610.g014

distance of more than 95% of 1st and 2nd degree relatives and more than 60% of relatives up to 3rd degree. At the same time, *PREPARE* has a higher PPV up to pedigrees of 4 generations. In the 5th generation it gets a lower PPV than the other methods, but this disadvantage is not meaningful, since the sensitivity of these methods in the 5th generation is very low. *PREPARE* gives better quality of results for larger populations, which is natural, since they tend to form simpler pedigrees with less multi-relationships between families, and less inbred families.

Considering *RMSIBDE* scores, *PREPARE* gets much better scores than the second best method, and is close to the optimal score, especially for larger populations. *IPED* gets worse *RMSIBDE* scores than *CIP/COP* as a result of its practical tendency to over-predict inbreeding, which we observed during our experiments. An important feature of *PREPARE*'s score is that it is non-increasing in the number of generations, similarly to the optimal score. In contrast, we do not see this behavior in other methods. Interestingly, the optimal scores decrease as the population size increases. We attribute this mainly to the increasing proportion of unrelated pairs in larger populations, which are easier to predict.

3.4 The Effect of Population Expansion on the Success of Pedigree Reconstruction

The simplified Wright-Fisher model that was used in pedigree reconstruction methods up to this day assumes a constant population size. Real populations sizes are obviously not constant, and it is known that population bottlenecks and expansion affect the IBD distribution in the population. We have conducted an experiment to test the effect of population size shifts on the distribution of chosen IBD features, and as a consequence on the quality of the resulting pedigree. We have run the Wright-Fisher simulation with changing initial population sizes of 100,200,300,400,500 and fixed the final population size at 500. By looking at the distribution of IBD features between all pairs of individuals, it is clear to see that the number of IBD segments and the mean IBD segment length have an inverse relationship with the initial population size. This corresponds to a higher proportion of relatives in the populations with smaller initial size. We have found that populations that grow from 100 to 500 individuals in five generations have similar IBD feature distributions to populations with constant population size of size 200. Interestingly the quality of the resulting pedigree of these populations remains unchanged when the initial population size is gradually decreased from 500 to 200. Only at initial size of 100 does the quality decrease. Sensitivity levels for initial population size of 100 are 0.96,0.75, and 0.54 for 2,3 and 4 generations. The largest decrease is for 3-generation pedigrees where the sensitivity is decreased by 10% on average. The PPV remains above 0.95 for generation 2,3 but is decreased from 0.85 to 0.71 in generation 4.

3.5 Comparing *PREPARE* and Competing Methods on Polygamous Simulations

To asses the quality of *PREPARE* on polygamous populations, we simulated polygamous populations of sizes 200 and 1000 with the Wright-Fisher model. In the simulated populations 33% of the siblings are half-siblings on average. Details regarding the execution of previous methods are the same as in section 3.3. *PREPARE* was run with the polygamous mode. The results are summarized in Fig. 11. Once again *PREPARE* is generally superior in terms of sensitivity, PPV and *RMSEIBD*. A notable exception is *IPED*'s relatively high sensitivity in generations 4 and 5 in smaller population sizes (200). Note however that this sensitivity comes at the cost of very low PPV and very high *RMSEIBD* in these generations. The *RMSEIBD* of *IPED* is not shown in the graph since it is out of the charts, getting as high as 1500 Mbp. This result suggests that *IPED* has a strong tendency to over-predict relationships in small polygamous populations.

Similarly to the monogamous case, *PREPARE* achieves higher performance on larger, and as a result, more simply related populations. For a population size of 1000, *PREPARE* is able to build a polygamous pedigree which correctly predicts the minimal meiosis distance of more than 97% of 1st degree relatives and more than 80% of 2nd degree relatives while maintaining a PPV greater than 80%. Polygamous populations pose a much greater challenge for pedigree reconstruction, and the performance is decreased in comparison to monogamous populations. According to our analysis, the difficulty in reconstructing polygamous pedigrees stems from the fact that the IBD feature distributions for the range of possible polygamous relationships have greater overlap than in monogamous relationships (See Fig. 12).

3.6 Reconstructing Realistically Simulated HapMap Descending Population

We test the performance of *PREPARE* on populations produced by the polygamous, asynchronous forward simulator. We run the simulator for hundreds of simulation years, resulting in the mixing of the different generations, and reconstruct the last five generations. We use un-phased IBD segments, to account for the fact that our input is genotypes, and not haplotypes. As a necessary step, we aim to filter out cross-generation relationships, which are not currently modeled, by taking the genotypes from the youngest age stratum (Ages 0-20). We used the *CEU* and *YRI* HapMap genotypes as the founder population for our simulation. The results show a comparable success to the Wright-Fisher simulation, increasing our confidence that *PREPARE* can be run on real populations. All accuracy measures show a decrease in accuracy compared to the Wright-Fisher simulation results. This is expected due to the addition of several factors (as discussed above), which adds to the complexity of the analysis (see Fig. 13).

3.7 Application for the HapMap MEX Population

We next use *PREPARE* to reconstruct the historical pedigree for the HapMap MEX population. This population is of interest to us since it is known to contain several relatives, including a single 4-generation pedigree [5]. Age information is not publicly available for this dataset. Instead, we use known parent-offspring relationships to separate the population into three generations. The correct pedigree is not known, so we use previous relationship inference results by Stevens et al. to validate our results[18].

Running *PREPARE* on the parent generation of HapMap phaseII+III *MEX* genotypes, we are able to detect a single sibling relationship (NA19662,NA19685), three first-cousin relationships (NA19662,NA19664), (NA19664,NA19685), (NA19657,NA19786) and two second-cousin relationships (NA19657,NA19785), (NA19785,NA19786). We are able to reconstruct correctly the pedigree found by Kyriazopoulou et al. We do this fully automatically and without using the genotypes of the two known grandparents: (NA19662,NA19685) which makes the reconstruction a significantly harder task(see Fig. 14). Further more, all of the relationships inferred by *PREPARE* except (NA19785,NA19786) are confirmed by Stevens et al.[18]. (NA19657,NA19786) are inferred as Third degree instead of first cousins, and (NA19657,NA19785) as Unknown degree instead of second cousins.

Discussion

In this paper, we take a step towards making pedigree reconstruction from present living populations, a realistic objective.

By developing better quality assessment tools, we were able to come to the conclusion that our method reconstructs pedigrees with significantly higher quality then previous methods, and in comparable running times. *PREPARE* is the first method to our knowledge to address polygamy, and paternal/maternal relative partitioning. Although we succeed partitioning the relatives, there is no way to know which relatives are really related to the father, and which to the mother by considering autosomal data alone. We are not worried about this lack in specificity, as we do not strive to learn the ancestral genders. Instead, we are interested in inferring the pedigree structure, which provides the relatedness structure. Our graph framework, brings to the surface several ambiguous cases that cannot be solved without utilizing additional subtle information. For example, the assignment of a 3-clique (see Fig. 5-B) might be decided better by considering three-way IBD sharing. The chance of having triple IBD sharing diminishes much faster than the chance of pair-wise IBD sharing and limits the theoretical possibility to correctly reconstruct these cases in advanced generations. Reconstructing inbred relationships correctly remains an unmet challenge by all methods in the present. It seems that an approach to deal with inbreeding will need to utilize additional inbreeding imprints on the data, such as homozygosity levels and other IBD-features not used today. Additionally, current methods do not include inbreeding options in the hypothesis testing stage, which might lead to the wrong conclusions when inbreeding exists. Despite the above, our method is able to reconstruct high quality pedigrees by dealing correctly with the most frequently arising cases in randomly mating populations. We believe that improving the performance on such rare aspects will probably have a small impact on the pedigree quality. More importantly, in order to further improve the reconstruction quality of polygamous populations, it seems that a better set of IBD features needs to be found, with higher separating power between different relationship types. Theoretically, the size of a family can influence the scores of its founders since larger families will contribute more extant individuals to the score computation. Simulating populations with differing typical family sizes show little effect on the quality of reconstruction. The current *PREPARE* method can be applicable for real populations, with the setback that only a specific age-range must be taken as input, such that most inter-generation relationships will be excluded.

Acknowledgments

We would like to thank Bonnie-Kirkpatrick and Dan He for their aid in successful compilation and running of the *CIP/COP* and *IPED* tools. Additionally, we would like to thank Moshe Einhorn and Roni Vilenchick, who worked with us on a project on the pedigree-reconstruction subject, which was the starting point for this research.

Author Contributions

Conceived and designed the experiments: DST EH. Performed the experiments: DST. Analyzed the data: DST. Contributed reagents/materials/analysis tools: DST EH. Wrote the paper: DST EH. Software design,development and testing: DST.

References

1. Blouin MS (2003) DNA-based methods for pedigree reconstruction and kinship analysis in natural populations. Trends in Ecology & Evolution 18: 503–511.
2. Lin TH, Myers EW, Xing EP (2006) Interpreting anonymous DNA samples from mass disasters–probabilistic forensic inference using genetic markers. Bioinformatics (Oxford, England) 22: e298–306.
3. Vouillamoz JF, Grando MS (2006) Genealogy of wine grape cultivars: "Pinot" is related to "Syrah". Heredity 97: 102–10.
4. Thomas SC, HillWG (2000) Estimating quantitative genetic parameters using sibships reconstructed from marker data. Genetics 155: 1961–72.
5. Kyriazopoulou-panagiotopoulou S, Haghighi DK, Aerni SJ, Sundquist A, Bercovici S, et al. (2011) Reconstruction of genealogical relationships with applications to Phase III of HapMap. Bioinformatics 27: 333–341.
6. Huff CD, Witherspoon DJ, Simonson TS, Xing J, Watkins WS, et al. (2011) Maximum-likelihood estimation of recent shared ancestry (ERSA). Genome research 21: 768–74.
7. Thompson Ea (1976) Inference of genealogical structure. Social Science Information 15: 477–526.
8. Almudevar A (2003) A simulated annealing algorithm for maximum likelihood pedigree reconstruction. Theoretical Population Biology 63: 63–75.
9. McPeek MS, Sun L (2000) Statistical tests for detection of misspecified relationships by use of genome-screen data. American journal of human genetics 66: 1076–94.
10. Cussens J, Bartlett M, Jones EM, Sheehan Na (2013) Maximum likelihood pedigree reconstruction using integer linear programming. Genetic epidemiology 37: 69–83.
11. Kirkpatrick B, Li SC, Karp RM, Halperin E (2011) Pedigree reconstruction using identity by descent. Journal of computational biology a journal of computational molecular cell biology 18: 1481–93.
12. Thatte BD, Steel M (2008) Reconstructing pedigrees: a stochastic perspective. Journal of theoretical biology 251: 440–9.
13. Steel M, Hein J (2006) Reconstructing pedigrees: a combinatorial perspective. Journal of theoretical biology 240: 360–7.
14. He D, Wang Z, Han B (2013) IPED : Inheritance Path Based Pedigree Reconstruction Algorithm Using Genotype Data. Recomb : 75–87.
15. Witherspoon DJ, Huff CD, Zhang Y, Watkins WS, Simonson TS, et al. (2010) ERSA E stimation of R ecent S hared A ncestry by maximum likelihood modeling of pairwise Applications of relationship estimation :. Genome research 21: 768–74.
16. Hstad J (1996) Clique is hard to approximate within n(1-ε) : 627–636.
17. Gusev A, Lowe JK, Stoffel M, Daly M, Altshuler D, et al. (2008) Whole Population, Genomewide Mapping of Hidden Relatedness. Genome research 19: 1–39.
18. Stevens EL, Heckenberg G, Roberson EDO, Baugher JD, Downey TJ, et al. (2011) Inference of relationships in population data using identity-by-descent and identity-by-state. PLoS genetics 7: e1002287.